A-95-35
II-A-5

QA/QC Protocol, Version 2.0:
Applying DQOs and DQA to MACT Standard Setting Decisions for Combustion Turbines and Reciprocating Internal Combustion Engines

Final Draft

Submitted to
Emission Measurement Center
Source Characterization Group B
MD-19, EMAD/OAQPS
Research Triangle Park, N.C., 27711

Prepared by
Michael J. Messner
Work Assignment Leader
Hadley M. Wood
Research Triangle Institute

U.S. EPA Contract 68D40091
Work Assignment 96-09
RTI Task 91U-6342-109

September 30, 1996

TABLE OF CONTENTS

List of Figures

List of Tables

BACKGROUND

This report provides a follow-up to the Draft Version 1.0 QA/QC Protocol released on March 26, 1996, by OAQPS. More specifically, this report presents recommended approaches to determining whether or not to develop a Maximum Achievable Control Technology (MACT) standard for two types of combustion engines. This problem and recommended approaches are presented in Data Quality Objectives (DQO) Process format, a systematic planning and decision-making framework developed by EPA's Quality Assurance Division. Historically, a significant amount of resources have been expended in the development of MACT standards. Although thresholds have been set by EPA to determine whether a standard needs to be developed, the data gathered by EPA to evaluate whether or not emissions exceed the thresholds often present a puzzling picture. For example, EPA's threshold for a single Hazardous Air Pollutant (HAP) is 10 tons/yr. Clearly, if many sources are emitting over 10 tons/yr of a specific HAP, EPA would develop a MACT standard. However, ambiguity arises when the data reveal only one or two data points (out of a large population) that exceed the 10 ton/yr threshold. Should EPA expend the resources required to develop and enforce a new MACT standard based upon these one or two points? This report explores these types of questions and presents approaches that may assist OAQPS in determining how confident they can be about MACT standard-setting decisions.

The proposed approach taken by the Emission Measurement Center (EMC) and Emission Standards Division (ESD) for reviewing both Reciprocating Internal Combustion Engine (RICE) and Combustion Turbine (CT) MACT emission data reports is based on a management decision to provide better reviews in less time and at a lower cost. In the past, all reports were submitted to EMC, who reviewed all the data in each test report. The objective of these reviews was to ensure that there was sufficient documentation and confirm, through independent QA/QC review, that the data were given one of 4 quality ratings:

1. Documentation is complete and data are of sufficient quality to be used to support standard development.

2. Data appear to be acceptable with the following minor deviations noted.

3. Data appear to be of sufficient quality but some documentation is missing. Recommend that additional documentation be obtained to upgrade to next level, or failing this, that the data be used to support other data on which to set the standard.

4. Data are obviously flawed; recommend that these data not be used for standard-setting purposes.

An important objective of EPA standard setting is that all data that are used for decision making purposes are of known high quality. In the past, EMC supported this objective by performing thorough, time-consuming reviews of data reports for each pollutant and each test method. ESD then used these quality ratings to determine which data sets they would base their final recommendation for a standard upon.

This previous approach was conservative and presented two major disadvantages. First, some valid data were excluded from further use because of insufficient documentation or poor report preparation. Second, a large amount of resources (including staff time and contractor dollars) were being expended to review data that were eventually discarded anyhow.

In this revised procedure, ESD staff must define the decision the data are being considered for, identify the key data that drive the decision, and define the level-of-certainty with which the decision must be made. EMC will then use ESD's database and their level of certainty specifications to judge how extensive the review must be.

For RICE and CT MACT, the decision currently being considered is "Does EPA need to develop a MACT standard?" To facilitate this ESD decision process, RTI has developed two possible decision rules to define the outcomes and the effects of making the "wrong" decision at this time in the regulatory process. The decision rules are presented below in DQO format (Step 5). The uncertainty that ESD is able to tolerate is an output of the Process (Step 6), and will help define how throughly EMC should review each subject report.

1.0 DATA QUALITY OBJECTIVES

Step 1. State the Problem

ESD is currently deciding if MACT standards for CTs and RICEs are required. Because of the wealth of test data available (approximately 160 CT and 1100 RICE pollutant test sets), traditional test report review is too expensive. An alternative test report review procedure is needed that reduces needed EMC staff resources to a manageable level.

Step 2. Identify the Decision

Upon preliminary review of the available data sets, the majority of the data points were found to be significantly below the MACT standard-setting threshold of 10 tons/yr. However, EPA was not able to make a "do not develop a MACT standard" determination for some of the data sets because one or two data points (within those data sets) were above the 10 ton/yr threshold. This finding allowed the problem statement to be simplified to the following decision statement: "How can EPA determine whether or not to develop a MACT standard when a single data point is above the MACT standard-setting threshold?" Two alternative actions arise from this problem:

- EPA can decide that the high data point is an outlier, throw that value out, and not develop a MACT standard; or
- EPA can decide that the high value is not an outlier, and proceed with the development of a MACT standard.

The decision statement and alternative actions specified above can be combined into the following statement:

If a data set contains a single point above the MACT threshold, EPA should determine whether the point is representative of the population as a whole and then proceed with development of a MACT standard only if the data point is determined to be significant.

If the decision is made to proceed with a MACT standard, then additional review of data critical to setting the level of the standard will be needed. This will likely be a different subset of the existing reports or new data.

Other, more specific, decision statements may arise in this DQO Process application. For example, two MACT thresholds exist under the Clean Air Act definition: (1) 10 tons/year per facility for a single hazardous air pollutant (HAP), or (2) 25 tons/year per facility total HAPs. This application of the DQO Process will address both thresholds.

Step 3. Identify Inputs to the Decision

- Existing data (in tons/yr) for individual HAPs and for total HAPs. ESD will lead on assembling the data.
- Potential for "high" and "low" bias in reported emission levels. EMC will lead on this

effort.

- Clean Air Act definition of "major source" as more than 10 tons/yr of any one HAP and more than 25 tons/yr of total HAPs.
- Information on *existing* control technologies and the *potential* to control HAP emissions. ESD will gather this information.
- Potential for co-location of multiple sources at the same facility.

Step 4. Define the Study Boundaries

The decision will apply to existing and future (new) Combustion Turbine and RICE HAP emissions. The decision will be made based upon assessment of the emissions data and will apply:

(A) until the Clean Air Act criteria for MACT development are satisfied in the future (if ever), if the decision is "don't develop the standard."

(B) until data to support standard setting are acquired, a thorough analysis of these data is conducted, and a new MACT standard is set, if the decision is to "develop the standard."

The geographic boundaries to which this decision applies are all CT and RICE systems in the United States. Although the data available for this study were gathered primarily from California, this study assumes that the data available for this assessment are representative of all CT and RICE systems in the U.S.

Step 5. Develop the Decision Rule

Two decision rules may be applied:

1. For each category of engine (CT and RICE), develop a regulatory standard for all the engines in that category if any one engine sampled is found to emit more than 10 tons per year of a single HAP or more than 25 tons per year of total HAPs.

2. For each category of engine (CT and RICE), develop a regulatory standard for all the engines in that category in the U.S. if more than X% (value to be determined) of the engines using that fuel are estimated to exceed the 10 ton/yr limit for a single HAP or 25 tons/yr of total HAPs.

The first decision rule is driven directly by the data. MACT standard setting would proceed if a single major source is found among the sources that were tested. The second decision rule is driven by what the data reveal about the larger population which they represent. MACT standard setting proceeds if the data show that a significant percentage of that population are major sources.

Step 6. Specify Tolerable Limits on Decision Errors

Given the first decision rule (1), two types of decision errors could occur when testing outliers:

(1A.) EPA could determine that the data point(s) above the MACT threshold are outliers and should be discarded and a MACT standard should not be developed *when, in actuality, the high data point(s) represent actual device emissions and a standard should have been developed.*

(1B.) EPA could determine that the data point(s) above the MACT threshold are not outliers and go ahead with the development of a MACT standard *when, in actuality, the high data point(s) do not represent actual emissions and the device they characterize is not a major source.*

Given the second decision rule (2), two types of decision errors could occur when testing outliers:

(2A.) There are actually more than X% major sources in the category and EPA decides to not develop a MACT standard.

(2B.) There are actually fewer than X% major sources in the category and EPA decides to go ahead with the MACT standard.

Consequences of error

Although this application of the DQO Process identifies two decision rules, the consequences of error under each decision rule are similar.

The consequences of decision error A (1A and 2A) may be significant. If EPA were to wrongly decide not to regulate using a MACT standard, human health and the environment would continue to be placed at risk. If the error is identified as a result of future study, the Agency could suffer some loss of credibility and may be criticized for its selective review and discarding of "high" results.

The consequences of decision error B (1B and 2B) may also be significant. An incorrect decision by EPA to develop a MACT standard would result in wasted resources, loss of Agency credibility, and potential economic impacts such as plant closings. A potential positive effect could be some marginal improvement of human health and environment.

Step 7. Optimize the Design for Obtaining Data

This step of the DQO Process is not relevant to this application because the data obtained for this study were generated by industry and offered to EPA for review.

2.0 DATA QUALITY ASSESSMENT OF CT/RICE DATA

Overview

The analysis of CT and RICE emissions data followed the Data Quality Assessment (DQA) Process, as described in EPA QA/G-9. The sections of this report are the steps of the DQA Process.

Step 1. Review of the DQOs

To review the DQOs, the data assessor revisits the outputs of the DQO Process to ensure that they still apply after data collection. As the data were acquired from industry, Step 7, Optimize the Design for Obtaining Data, was not relevant to this DQO application. To the extent practical, however, this DQA will assess uncertainties so that EPA will be able to make an informed decision about whether to pursue MACT standard setting for the various sets of Combustion Turbines (CTs) and Reciprocating Internal Combustion Engines (RICEs). The critical output of the DQO Process is Step 5, Develop the Decision Rule. The decision rule consists of two alternatives as described below.

1. For each category of engine (CT and RICE), develop a regulatory standard for all the engines in that category if any one engine is found to emit more than 10 tons per year of a single HAP or more than 25 tons per year of total HAPs.

2. For each category of engine (CT and RICE), develop a regulatory standard for all the engines in that category if more than X% (value to be determined) of the engines using that fuel are estimated to exceed the 10 ton/yr limit for a single HAP or 25 tons/yr of total HAPs.

This DQA will address both decision rules, but EPA must determine which rule will be applied to the final decision about whether or not to develop a MACT standard for each engine. The first option would be a simple decision based directly upon the collected data. The second option would require a statistical estimation of the population that the collected data represent. For this DQA, RTI used X=5% for decision rule Option 2.

Step 2. Preliminary Data Review

Next, the data assessor conducts a review of the data to determine obvious patterns, relationships, or potential anomalies. This review consists of a general overview of the data sets and an initial analysis.

Overview of the Data Sets

The emissions data are divided into two groups based on the engine type, CT and RICE.

The CT data consists of five fields: engine fuel type, pollutant emitted, identification (ID) number of the location of emissions, pounds of pollutant emitted per hour of engine operation, and tons of pollutant emitted per year of engine operation. The data consists of 136 records with five fuel types, 32 ID numbers, and 26 pollutants. Two of the locations exceeded the emission threshold of 10 tons per year for an individual pollutant with values of 11.5194 and 127.05504 tons of formaldehyde per year respectively. Only one location (a natural gas burning turbine) exceeded the total pollutant emission threshold of 25 tons per year with 127.0869 tons of emissions. This same facility was one of the two facilities that exceeded the 10 ton/yr threshold for a single HAP (formaldehyde).

The RICE data contains the same five fields as the CT data. However, the RICE data set is much larger, containing 1071 records with 5 fuel types, 90 ID numbers, and 42 pollutants. Only location ID number 10 exceeds the yearly single pollutant threshold with 34.5144 tons per year of formaldehyde emitted. This is also the only location to exceed the annual total pollutant standard with 34.6165854 tons per year of emissions. Over 99% of the total HAPs for this single location was comprised of a single HAP (formaldehyde).

Initial Analysis

The initial analysis of the data was conducted on each category of engine. Within each category, only formaldehyde exceeded the emissions thresholds. Therefore, only formaldehyde emissions data were reviewed. For each engine category, the data were also subdivided by fuel type. For the CT engines, natural gas and distillate oil were the predominant fuel types used. For the RICEs that were tested, natural gas, digester gas, and diesel fuel were the major fuel types used. This analysis assumes that the population of fuel types is adequately represented in the sample population. Therefore, only data from these fuel types were reviewed. In addition, it assumes that each data set represents "normal operation" of that facility. From these two assumptions, we can assume that the data sets represent a randomly gathered subset of the population of engines in the U.S. Consequently, the analysis was performed on the following subsets of the data: (1) formaldehyde emissions from CT engines using natural gas fuel, (2) formaldehyde emissions from CT engines using distillate oil fuel, (3) formaldehyde emissions from RICEs using natural gas, (4) formaldehyde emissions from RICEs using diesel fuel, and (5) formaldehyde emissions from RICEs using digester gas fuel.

The initial analysis entailed calculating statistical quantities, determining the distribution of the data, generating graphs, and testing for data outliers for the above 5 data sets. This analysis assumes that the fuel types are adequately represented by this sample population. In addition, it is assumed that each data set represents "normal" operation of that facility and that we can assume that the data sets represent a randomly gathered subset of the entire population. These assumptions should be independently confirmed. The statistical quantities calculated were the sample size, minimum, maximum, mean, median, variance, and the standard deviation. The distribution of each data set (normal versus lognormal distribution) was evaluated using the Shapiro-Wilk test (Table 1). All the formaldehyde data appear to be lognormally distributed except for the RICE using digester gas data, for which the normal distribution provided the best fit. The four data sets that were determined to be lognormally distributed were log transformed. Graphs of the data their log transformations were produced (Figures 1-10). Finally, Dixon's outlier test was used in a search for outliers. None of the data sets were determined to contain outliers. Note: Log-transformed statistical quantities (min., max., mean, median, variance, and standard deviation) are presented in Table 1 for all the data that appeared to be lognormally distributed.

Step 3. Select the Statistical Test

The statistical test used to evaluate the data for determining if regulation will be needed depends on the decision rule selected. If Option 1 is selected, no statistical test will be required. The data assessor would need only to determine whether any value in the data set exceeds the threshold. Option 2 would require that the data assessor estimate the percentage of the underlying population that exceeds the threshold. This involves estimating the percentage of the population that exceeds ln(10) for log-transformed data. Numerical integration techniques were used to produce a 90% confidence interval for the percentile estimates.

Step 4. Verify the Assumptions of the Statistical Test

Distributional assumptions were tested using Shapiro-Wilk test (See Step 2, above) and probability plots were produced for additional verification that the distributional assumptions were reasonable. These plots support the Shapiro-Wilk test results.

Step 5. Draw Conclusions

Table 1 provides the estimated quantities, together with indications of their uncertainty. The five sections that follow address the five individual data sets:

2.1. CT - Natural Gas
2.2. CT - Distillate
2.3. RICE - Natural Gas
2.4. RICE - Diesel Fuel
2.5. RICE - Digester Gas

2.1 CT - Natural Gas (CT_Ngas)- Annual Formaldehyde Emissions

Distributional Assumptions and Outlier Test

Figures 1 and 2 show the poor fit of the normal distribution and good fit of the lognormal distribution, respectively. The Y-intercept and slope of Figure 2 represent the mean and standard deviation of the transformed data set. The extreme points of the ln-transformed data were tested using the Dixon Criteria. No outliers were identified (Table 1).

Estimation of Percentiles

The probability that a single device (selected at random from the population of interest) will have formaldehyde emissions greater than 10 tons/yr depends on the estimated mean (Y_bar), standard deviation (s_Y), and degrees of freedom (ν) derived from the appropriate log-transformed data set. The probability of a single device exceeding 10 tons/yr is given by the Student's t distribution (with ν degrees of freedom):

$$Pr\{t_\nu >[\ln(10) - Y\_bar] / s_Y\}$$

Where:

t_ν = Student's t-value with ν degrees of freedom
Y_Bar = estimated mean of the data set
S_y = standard deviation

The mean and standard deviation of ln-transformed annual formaldehyde emission estimates for natural gas fired combustion turbines (CT_Ngas) are -1.042 and 2.417, respectively (Table 1). For this group of turbines, then, the percentile represented by 10 tons/yr is given by:

$$Pr\{t_\nu > [\ln(10) - Y\_bar] / s_Y\} = Pr\{t_{18} > 1.384\} = 0.092$$

This value estimates the probability that a single turbine, selected at random from the larger population of combustion turbines, will produce greater than 10 tons/yr formaldehyde. The value 10 tons/yr was found to be at the 91.8th percentile of the distribution, indicating that 9.2% of the population of CTs fueled by natural gas are believed to be major sources. *In simpler terms, this value estimates that 9.2% of the CTs in this country that run on natural gas emit at least 10 tons of formaldehyde per year.*

The second step of this assessment involves estimation of the 95th percentiles of the distribution. The 95th percentile for each set was found as the value of Y that solved the following:

$$Pr\{t_\nu > [\ln(Y) - Y\_bar] / s_Y\} = 0.05$$

Where:

t_ν = Student's t-value with ν degrees of freedom
Y_Bar = estimated mean of the data set
S_y = standard deviation

For CT_Ngas, the value of Y solving the above is 3.1. Transformed back to measurement units, the 95th percentile is 22.4 tons/yr (Table 1). *In simpler terms, this value estimates that 5% of the CTs in this country that run on natural gas emit at least 22.4 tons of formaldehyde per year.*

The final step of this assessment involves derivation of the confidence limits for the 95th percentile estimates. For CT_Ngas, the 90% confidence limits were estimated to be 1.818 and 4.695, which transform to 6.16 and 109 tons/yr, respectively. *These values mean that the true 95th percentile is expected to lie between 6.16 and 109 tons per year, which implies that the assessors cannot be confident that less than 5% of the CTs are major sources.*

Removal of High Values from CT Natural Gas Data Set

At the request of EPA, the highest value was removed to ascertain its impact on the findings. The objective of this exercise is to test whether the removal of the highest data point does affected the determination of whether to regulate or not. If, in fact, the removal of this data point had no impact on EPA's decision about regulation, then extensive QA review of that data would be superfluous. The statistical quantities applicable to this log-transformed data set are presented as CT_Ngas2 on Table 1. Removal of this highest point (not shown in Figure 2) resulted in a 95th percentile estimate of 8.0 tons of formaldehyde per year. However, the upper confidence limit for this percentile was 32.9 tons/yr, well above the 10 ton/yr limit. Therefore, removal of the highest data point from this set would not allow the assessor to reject the null hypothesis that the 95th percentile was equal to 10 tons/yr. If EPA decided to use this decision rule, EMC review of this data set would be unnecessary, since a decision to regulate would be made regardless of whether the high data point was included or excluded from the data set. Next, the two highest values were removed from the CT_Ngas data set (not shown in Figure 2). The statistical quantities for this log-transformed data set are presented as CT_Ngas3 on Table 1. Although this resulted in a 95th percentile estimate of 4.69, the upper confidence limit was 17.81 tons/yr, which was still above the 10 ton/yr limit. Therefore, removal of both of these values would not allow for the EPA to reject the null for Option 2, which is that at least 5% of these turbines are major sources. However, removal of both of the high values will allow EPA to reject the null under Option 1, since all the remaining values (after removal of the two high data points) are well below the 10 ton/yr threshold. *If EPA were to decide either not to remove any of the high data points or if they were to decide to remove only the highest data point (CT_Ngas and CT_Ngas2), both decision rules (Option 1 and Option 2) fail and EPA would be required to develop a MACT standard for formaldehyde from natural gas-burning Combustion Turbines. If they were to decide to remove both of the data points that exceeded the 10 ton/yr threshold (CT_Ngas3), they would still be required to develop a MACT standard under Option 2. However, they would not be required to develop a MACT standard under Option 1 for CT_Ngas3. Hence, only under decision rule Option 1 with both the high data points removed would any QA review be necessary. For all other cases (Option 2, no points removed; Option 2, one point removed; Option 2, two points removed; Option 1, no points removed; and Option 1, one point removed), no QA review would be necessary.*

The EMC QA review should furthermore focus on technical reasons why any reported data are biased high and should be done separately starting with the highest reported value. Unless reasons are found to remove that data point, then there is no reason to review the next highest reported value. This has effectively reduced the EMC report work load from 26 reports of varying combinations of compounds to 1 or 2 reports of formaldehyde only. Additional reviews of different subsets of data may be necessary to answer other questions.

Figure 1. CT Natural Gas - Normal Probability Plot

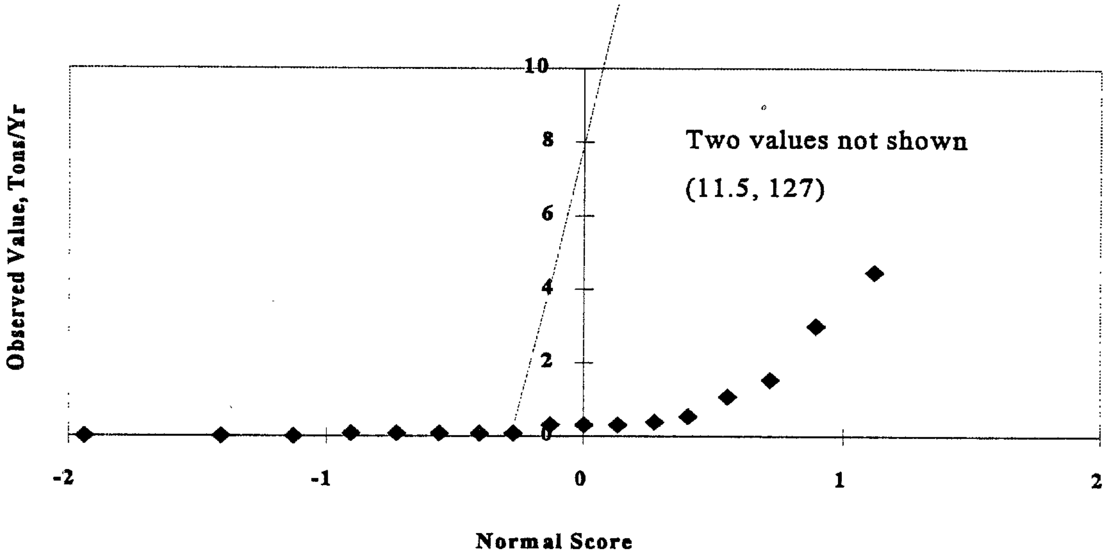
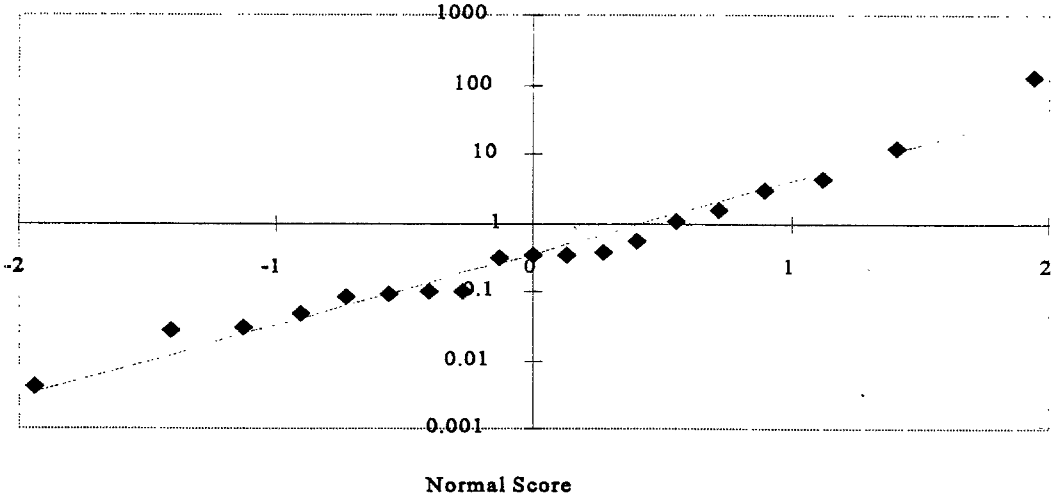


Figure 2. CT Natural Gas Lognormal Probability Plot



2.2 CT - Distillate Fuel (CT_Dist)- Annual Formaldehyde Emissions

Distributional Assumptions and Outlier Test

Figures 3 and 4 show the poor fit of the normal distribution and good fit of the lognormal distribution, respectively. The Y-intercept and slope of Figure 4 represent the mean and standard deviation of the transformed data set. The extreme points of the ln-transformed data were tested using the Dixon Criteria. No outliers were identified (Table 1).

Estimation of Percentiles

The probability that a single device (selected at random from the population of interest) will have formaldehyde emissions greater than 10 tons/yr depends on the estimated mean (Y_bar), standard deviation (s_Y), and degrees of freedom (ν) derived from the appropriate log-transformed data set. The probability of a single device exceeding 10 tons/yr is given by the Student's t distribution(with ν degrees of freedom):

$$\Pr\{t_\nu > [\ln(10) - Y\_bar] / s_Y\}$$

Where:

t_ν = Student's t-value with ν degrees of freedom
Y_Bar = estimated mean of the data set
S_y = standard deviation

The mean and standard deviation of ln-transformed annual formaldehyde emission estimates for natural gas-fired combustion turbines (CT_Dist) are -0.679 and 1.377, respectively (Table 1). For this group of turbines, then, the percentile represented by 10 tons/yr is given by:

$$\Pr\{t_7 > [\ln(10) - Y\_bar] / s_Y\} = 0.034$$

This value estimates the probability that a single turbine, selected at random from the larger population of combustion turbines, will produce in excess of 10 tons/yr formaldehyde. The value 10 tons/yr was found to be at the 96.6th percentile of the distribution, indicating that 3.4% of the population of CTs fueled by distillate oil are believed to be major sources. *In simpler terms, this value estimates that 3.4% of the CTs in this country that run on distillate oil emit at least 10 tons of formaldehyde per year.*

The second step of this assessment involves estimation of the 95th percentiles of the distribution. The 95th percentile for each set was found as the value of Y that solved the following:

$$\Pr\{t_\nu > [\ln(Y) - Y\_bar] / s_Y\} = 0.05$$

Where:

t_v = Student's t-value with ν degrees of freedom
Y_Bar = estimated mean of the data set
S_y = standard deviation

For CT_Dist, the value of Y solving the above is 1.87. Transformed back to measurement units, the 95th percentile is 6.49 tons/yr (Table 1). *In simpler terms, this value estimates that 5% of the CTs in this country that run on distillate oil emit at least 6.49 tons of formaldehyde per year.*

The final step of this assessment involves derivation of the confidence limits for the percentile estimates. For CT_Dist, the 90% confidence limits were estimated to be 0.643 and 3.708, which transform to 1.90 and 40.8 tons/yr, respectively. *These values mean that the true 95th percentile is expected to lie between 1.90 and 40.8 tons per year, which implies that the assessors cannot be confident that less than 5% of theses CTs are major sources.*

Therefore, although under decision rule Option 1 EPA would not be required to develop a MACT standard for distillate oil-burning CTs, EPA would not be able to reject the null hypothesis for Option 2 and would, therefore, be required to develop a MACT standard.

Figure 3. CT_Distillate - Normal Probability Plot

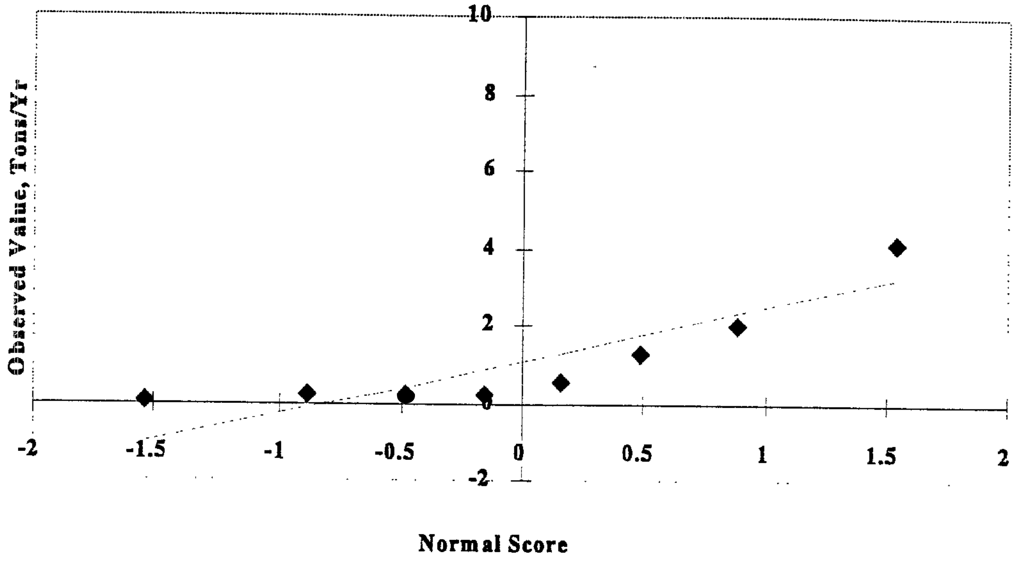
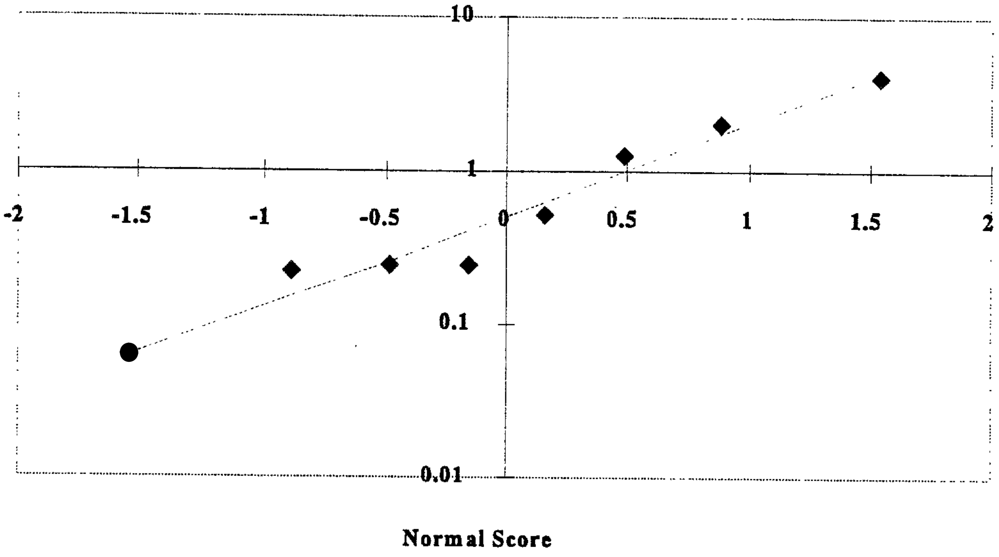


Figure 4. CT_Distillate - Lognormal Probability Plot



2.3 RICE - Natural Gas (R_NGas)- Annual Formaldehyde Emissions

Distributional Assumptions and Outlier Test

Several zeros were reported in the natural gas-fueled RICE data set. The normal and lognormal probability plots are represented as Figures 5 and 6, respectively. Clearly, normality is not a valid assumption. Note what appears to be a breakpoint on the lognormal plot (Figure 6) around 0.02 tons/yr. The data assessor decided to censor the values below 0.02 tons/yr in order to provide parameter for characterizing the upper tail of the distribution estimates. A mean of -1.135 and standard deviation of 2.041 maximized the likelihood function for ln-transformed data(R_Ngas, Table 1). The extreme points of the ln-transformed data were tested using the Dixon Criteria. No outliers were identified.

Estimation of Percentile

The probability that a single device (selected at random from the population of interest) will have formaldehyde emissions greater than 10 tons/yr depends on the estimated mean (Y_bar), standard deviation (s_Y), and degrees of freedom (ν) derived from the appropriate log-transformed data set. The probability of a single device exceeding 10 tons/yr is given by the Student's t distribution (with ν degrees of freedom):

$$\Pr\{t_\nu > [\ln(10) - Y\_bar] / s_Y\}$$

Where:

t_ν = Student's t-value with ν degrees of freedom
Y_Bar = estimated mean of the data set
S_y = standard deviation

For this group of engines, then, the percentile represented by 10 tons/yr is given by:

$$\Pr\{t_{51} > [\ln(10) - Y\_bar] / s_Y\} = .049$$

This value estimates the probability that a single engine, selected at random from the larger population of RICEs, will produce in excess of 10 tons/yr formaldehyde. The value 10 tons/yr was found to be at the 95.1th percentile of the distribution, indicating that 4.9% of the population of RICEs fueled by natural gas are believed to be major sources. *In simpler terms, this value estimates that 4.9% of the RICEs in this country that run on natural gas emit at least 10 tons of formaldehyde per year.*

The second step of this assessment involves estimation of the 95th percentiles of the distribution. The 95th percentile for each set was found as the value of Y that solved the following:

$$\Pr\{t_\nu > [\ln(Y) - Y\_bar] / s_Y\} = 0.05$$

Where:

t_v = Student's t-value with v degrees of freedom
Y_Bar = estimated mean of the data set
S_y = standard deviation

For R_Ngas, the value of Y solving the above is 2.27. Transformed back to measurement units, the 95th percentile is 9.708 tons/yr (Table 1). *In simpler terms, this value estimates that 5% of the RICEs in this country that run on natural gas emit at least 9.71 tons of formaldehyde per year.*

The final step of this assessment involves derivation of the confidence limits for the 95th percentile estimates. For R_Ngas, the 90% confidence limits were estimated to be 1.557 and 3.010, which transform to 4.75 and 20.3 tons/yr, respectively. *These values mean that the true 95th percentile is expected to lie between 4.75 and 20.3 tons per year, which implies that the data assessors cannot be confident that less than 5% of these RICEs are major sources.*

Under both decision rules (Option 1 and Option 2) EPA is not able to reject the null hypothesis and would, therefore, be required to develop a MACT standard.

Removal of High Values from RICE Natural Gas Data Set

At the request of EPA, the highest value was removed to ascertain its impact on the findings. The statistical quantities applicable to this log-transformed data set are presented as R_Ngas2 on Table 1. Removal of this highest point (not shown in Figure 6) resulted in a 95th percentile estimate of 3.99 tons of formaldehyde per year. The upper confidence limit was 7.28 tons/yr, which falls below the 10 ton/yr limit. Therefore, removal of the highest data point from this set would allow the assessor to reject the null hypothesis that the 95th percentile of the distribution is equal to 10 tons/yr. *Since both decision rules (Option 1 and Option 2) are rejected with the removal of the highest data point in the R_Ngas data set, EPA would not be required to develop a MACT standard for formaldehyde from natural gas-burning RICEs if they chose this data assessment approach.* Hence, an EMC QA review of this data point is important. This review should focus on technical reasons why the reported data could be biased high by a factor of 5 to 10, since any small change would not affect the EPA decision to proceed with standard setting.

Figure 5. RICE Natural Gas - Normal Probability Plot

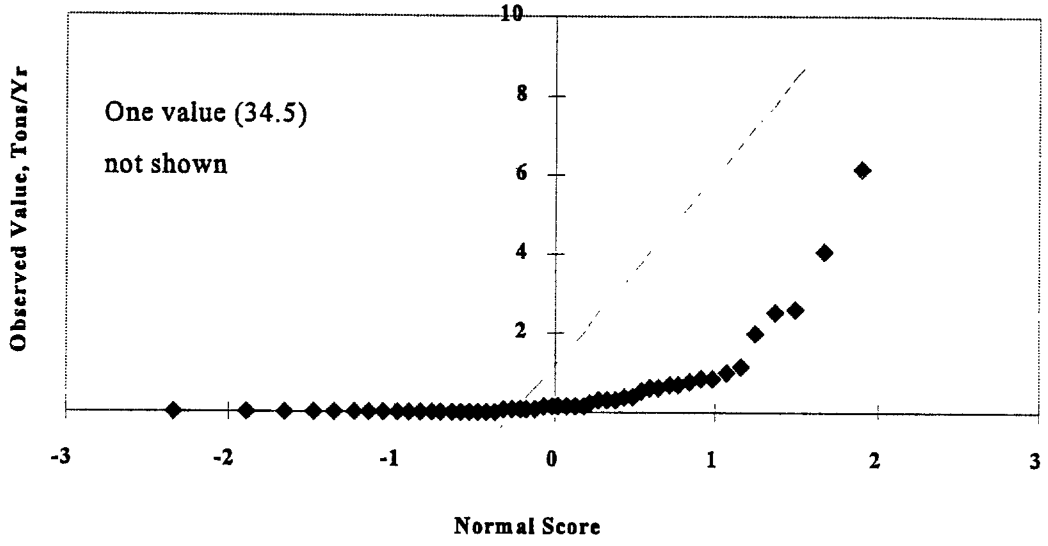
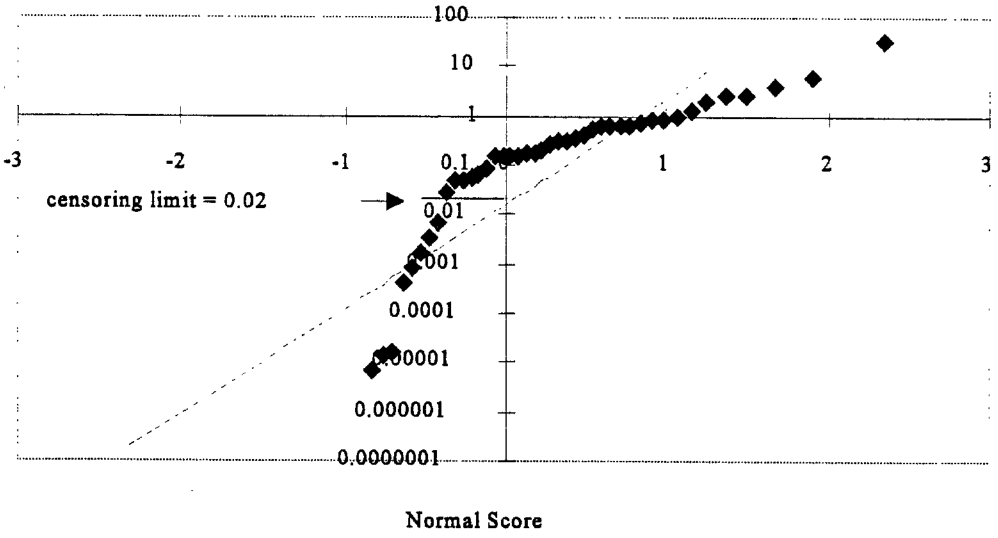


Figure 6. RICE Natural Gas - Lognormal Probability Plot



2.4 RICE - Diesel Fuel (R_DFuel)- Annual Formaldehyde Emissions

Distributional Assumptions and Outlier Test

Figures 7 and 8 show the poor fit of the normal distribution and good fit of the lognormal distribution, respectively. The Y-intercept and slope of Figure 8 represent the mean and standard deviation of the transformed data set. No outliers were identified in this data set.

Estimation of Percentiles

The probability that a single device (selected at random from the population of interest) will have formaldehyde emissions greater than 10 tons/yr depends on the estimated mean (Y_bar), standard deviation (s_Y), and degrees of freedom (ν) derived from the appropriate log-transformed data set. The probability of a single device exceeding 10 tons/yr is given by the Student's t distribution(with ν degrees of freedom):

$$\Pr\{t_\nu > [\ln(10) - Y\_bar] / s_Y\}$$

Where:

t_ν = Student's t-value with ν degrees of freedom
Y_Bar = estimated mean of the data set
S_y = standard deviation

The mean and standard deviation of ln-transformed annual formaldehyde emission estimates for diesel fueled RICEs (R_DFuel) are -5.275 and 1.874, respectively (Table 1). For this group of engines, then, the percentile represented by 10 tons/yr is given by:

$$\Pr\{t_{17} > [\ln(10) - Y\_bar] / s_Y\} < 0.001$$

This value estimates the probability that a single engine, selected at random from the larger population of RICEs, will produce in excess of 10 tons/yr formaldehyde. The value 10 tons/yr was found to be above the 99.9th percentile of the distribution, indicating that <0.1% of the population of RICEs fueled by diesel fuel are believed to be major sources. *In simpler terms, this value estimates that less than 0.1% of the RICEs in this country that run on diesel fuel emit at least 10 tons of formaldehyde per year.*

The second step of this assessment involves estimation of the 95th percentiles of the distribution. The 95th percentile was found as the value of Y that solved the following:

$$\Pr\{t_\nu > [\ln(Y) - Y\_bar] / s_Y\} = 0.05$$

Where:

t_v = Student's t-value with ν degrees of freedom
Y_Bar = estimated mean of the data set
S_y = standard deviation

For R_DFuel, the value of Y solving the above is -2.048. Transformed back to measurement units, the 95th percentile is 0.13 tons/yr (Table 1). *In simpler terms, this value estimates that the highest-polluting 5% of the RICEs in this country that run on diesel fuel emit only 0.13 tons of formaldehyde per year.*

The final step of this assessment involves derivation of the confidence limits for the 95th percentile estimates. For R_DFuel, the 90% confidence limits were estimated to be -3.112 and -0.674, which transform to 0.045 and 1.96 tons/yr, respectively. *These values mean that the true 95th percentile is expected to lie between 0.045 and 1.96 tons per year, which implies that the data assessors can be confident less than 5% of these RICEs are major sources.*

Therefore, under both decision rules (Option 1 and Option 2) EPA would not be required to develop a MACT standard for diesel fuel-burning RICEs. In this case, EMC QA review should focus only on technical reasons why formaldehyde test methods used would be expected to be biased low by a factor of 5 to 10 because this is what is needed to change the decision.

Figure 7. RICE Diesel Fuel- Normal Probability Plot

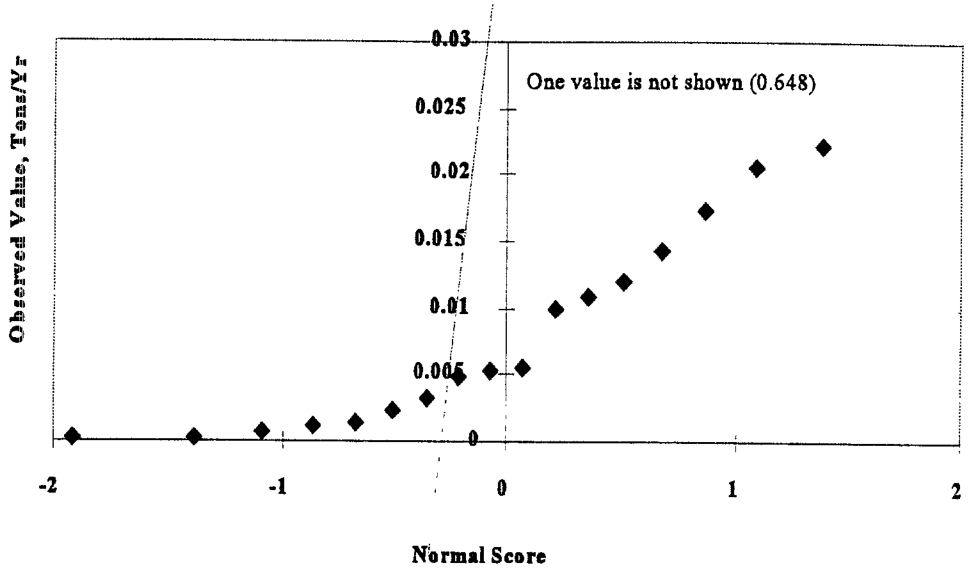
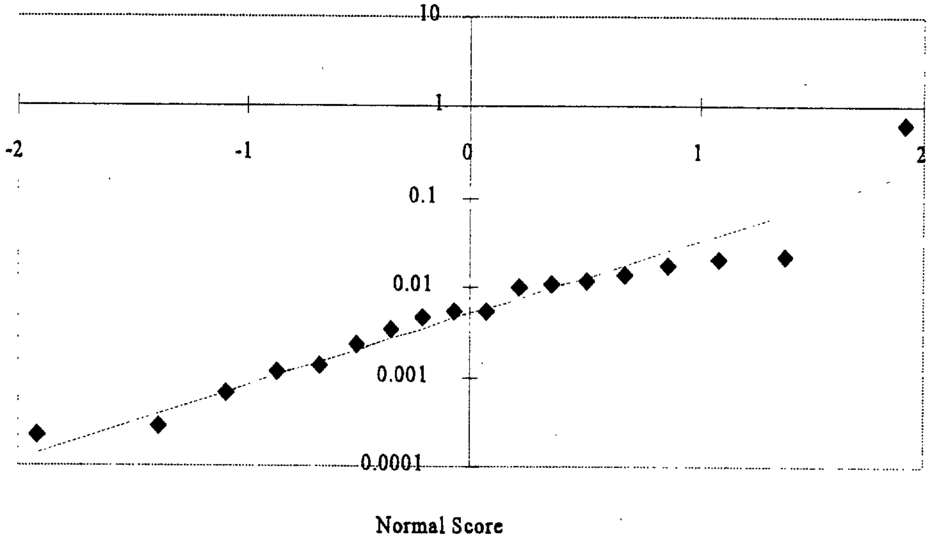


Figure 8. RICE Diesel Fuel- Lognormal Probability Plot



2.5 RICE - Digester Gas (R_DGas)

Distributional assumptions and outlier test

Figure 9 shows good fit of the normal distribution and Figure 10 shows the poor fit of the lognormal distribution. Note that this data set is the only one in this study that does not appear to be lognormally distributed. The Y-intercept and slope of Figure 9 represent the mean and standard deviation of the data set. No outliers were detected in this data set.

Estimation of Percentile

The probability that a single device (selected at random from the population of interest) will have formaldehyde emissions greater than 10 tons/yr depends on the estimated mean (X_bar), standard deviation (s_x), and degrees of freedom (ν) derived from the appropriate normally distributed data set. The probability of a single device exceeding 10 tons/yr is given by the probability that Student's t (with ν degrees of freedom) exceeds the following:

$$\Pr\{t_\nu > [10 - X\_bar] / s_x\}$$

Where:

t_ν = Student's t-value with ν degrees of freedom
X_Bar = estimated mean of the untransformed data set
S_x = standard deviation (untransformed)

The mean and standard deviation of annual formaldehyde emission estimates for digester gas-fueled RICEs (R_Dgas) are 0.022 and 0.020, respectively (Table 1). For this group of engines, then, the percentile represented by 10 tons/yr is given by:

$$\Pr\{t_{16} > [10 - X\_bar] / s_x\} < 0.001$$

This value estimates the probability that a single engine, selected at random from the larger population of RICEs, will produce in excess of 10 tons/yr formaldehyde. The value 10 tons/yr was found to be above the 99.9th percentile of the distribution, indicating that <0.1% of the population of RICEs fueled by digester gas are believed to be major sources. *In simpler terms, this value estimates that less than 0.1% of the RICEs in this country that run on digester gas emit at least 10 tons of formaldehyde per year.*

The second step of this assessment involves estimation of the 95th percentiles of the distribution. The 95th percentile for each set was found as the value of X that solved the following:

$$\Pr\{t_\nu > [X - X\_bar] / s_x\} = 0.05$$

Where:

t_v = Student's t-value with v degrees of freedom
X_Bar = estimated mean of the data set
S_x = standard deviation

For R_Dgas, the value of X solving the above is 0.057 tons/yr (Table 1). *In simpler terms, this value estimates that the highest-polluting 5% of the RICEs in this country that run on digester gas emit only 0.057 tons of formaldehyde per year.*

The third step of this assessment involves derivation of the confidence limits for the 95th percentile estimates. For R_Dgas, the 90% confidence limits were estimated to be 0.045 and 0.072 tons/yr. *These values mean that the true 95th percentile is expected to lie between 0.045 and 0.072 tons per year, which implies that the data assessors can be confident that less than 5% of these RICEs are major sources.*

Therefore, under both Decision Rules (Option 1 and Option 2) EPA would not be required to develop a MACT standard for digester gas-burning RICEs. In this case, EMC QA review should focus on technical reasons why formaldehyde test methods used would be expected to be biased low by a factor of 5 to 10, because this would be necessary before the EPA decision would be changed.

Figure 9. RICE Digester Gas - Normal Probability Plot

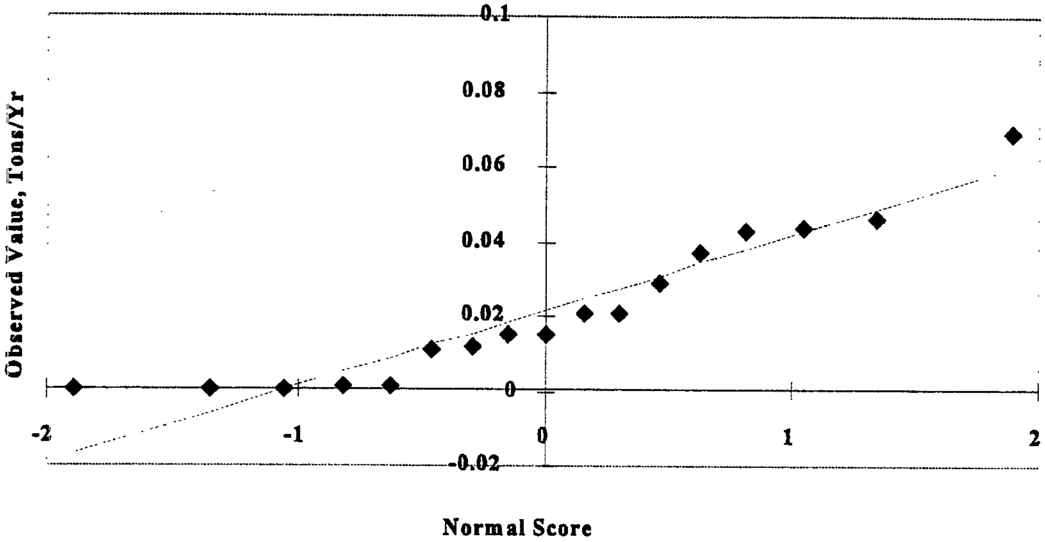


Figure 10. RICE Digester Gas - Lognormal Probability Plot

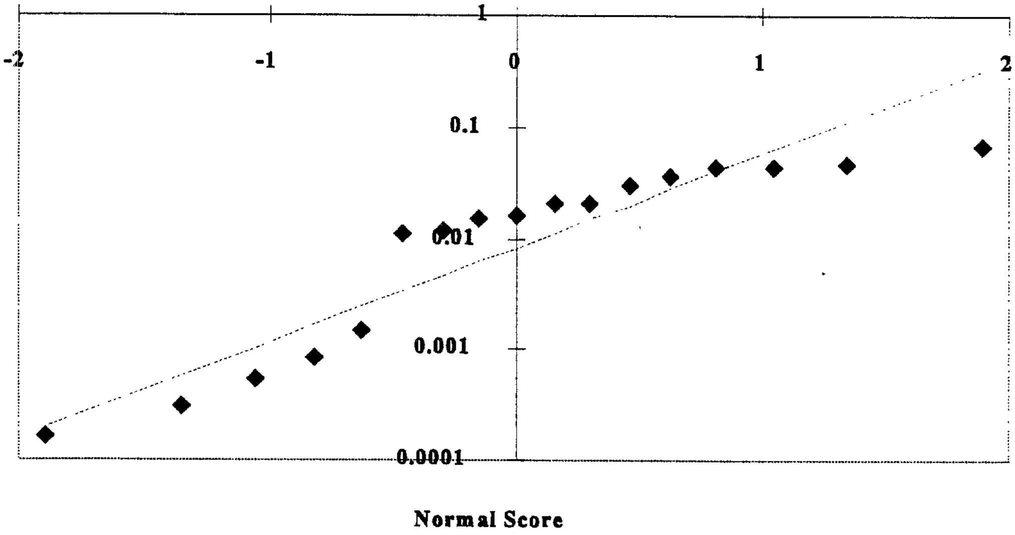


Table 1. Summary Statistics for CT and RICE Formaldehyde Emissions Data

	CT Ngas	CT Ngas2	CT Ngas3	CT Dist	R Ngas	R Ngas2	R Dfuel	R Dgas
N	19	18	17	8	52	51	18	17
minimum	-5.524	-5.524	-5.524	-2.785	0	0	-8.422	0.000
maximum	4.845	2.444	1.487	1.408	3.541	1.814	-0.433	0.069
mean	-1.042	-1.369	-1.593	-0.679	-1.135	-1.194	-5.275	0.022
med	-1.100	-1.152	-1.204	-1.031	-0.072	-0.142	-5.208	0.016
variance	5.841	4.033	3.323	1.897	4.166	2.384	3.513	0.000
std. dev.	2.417	2.008	1.823	1.377	2.041	1.544	1.874	0.020
distribution	lognormal	lognormal	lognormal	lognormal	lognormal	lognormal	lognormal	normal
outliers? (Dixon)	none	none	none	none	none	none	none	none
Pr{>10 T/yr}	0.092			0.034	0.049	0.014	<0.001	<0.001
95th %ile (Lower) ln transform	1.818	0.97	0.517	0.643	1.557	0.873	-3.112	---
95 %ile (expected) ln transform	3.109	2.08	1.546	1.87	2.272	1.384	-2.048	---
95 %ile (upper) ln transform	4.7	3.492	2.88	3.708	3.010	1.984	-0.6741	---
95 %ile (lower) T/yr.	6.6	2.64	1.68	1.90	4.745	2.394	0.045	0.045
95 %ile (expected) T/yr.	22.4	8.0	4.69	6.49	9.708	3.991	0.129	0.057
95 %ile (upper) T/yr.	110	32.9	17.81	40.8	20.29	7.28	0.510	0.072
Decision: Option 1?	Yes	Yes	No	No	Yes	No	No	No
Decision: Option 2?	Yes	Yes	Yes	Yes	Yes	No	No	No

3.0 ESTIMATING THE PROBABILITY THAT COLLOCATED UNITS WILL CONSTITUTE A MAJOR SOURCE

The analysis presented above presents an approach for estimating the probability that a single unit will constitute a major source of HAP emissions (and therefore require the development of a MACT standard). An additional question asked by EMC was: "What is the probability that two or more collocated units *together* constitute a major source?" This section of the report will address that question.

Approach for estimating the probabilities

The approach for estimating the probability that a single unit will be a major source (exceed 10 tons/yr of formaldehyde) is discussed above. Estimating the probability that collocated units constitute major sources required simulation (numerical integration proved impractical for more than 2 collocated devices). Simulations were conducted using @Risk for Excel. Probability distributions utilized Student's t. Included in the simulation were tests for one and two devices. The results for one and two devices were compared with the numerical integration results as a quality control check. The data assessor ran each simulation 10,000 times, providing reasonable accuracy in the probability estimates. For example, the probability that three CT_Ngas units will constitute a major source was determined to be 0.380 (3800 of 10,000 iterations produced trios that exceeded 10 tons/yr). The uncertainty of the estimate, expressed as standard error, is approximated by the square root of 3800 divided by 10,000:

Estimate = 0.380 Standard Error = 3800^0.5 / 10,000 = 0.006

Results

The results of these simulations for each of the data sets are presented in Tables 2-6, below. The column "Pr{Major Source}" indicates the probability that the number of devices specified in "Number of Collocated Devices" will constitute a major source.

Table 2. Simulation Parameters and Results for CT_Ngas

CT_Ngas	Number Collocated Devices	Pr{Major Source}

Y_bar = -1.04
s_Y = 2.42
ν = n - 1 = 18
iterations = 10,000
	1	0.092

	2	0.189
	3	0.283
	4	0.375
	5	0.469
	6	0.554

Table 3. Simulation Parameters and Results for CT_Dist

CT_Dist	Number Collocated Devices	Pr{Major Source}
Y_bar = -0.68	1	0.034
$s_Y = 1.38$	2	0.074
$\nu = n - 1 = 7$	3	0.122
iterations = 10,000	4	0.180
	5	0.248
	6	0.323

Table 4. Simulation Parameters and Results for R_Ngas

R_Ngas	Number Collocated Devices	Pr{Major Source}
Y_bar = -1.14	1	0.049
$s_Y = 2.041$	2	0.105
$\nu = n - 1 = 51$	3	0.167
iterations = 10,000	4	0.232
	5	0.302
	6	0.379

Table 5. Simulation Parameters and Results for R_Dfuel

R_Dfuel	Number Collocated Devices	Pr{Major Source}
Y_bar = -5.27	1	0.000
$s_Y = 1.87$	2	0.001
$\nu = n - 1 = 17$	3	0.002
iterations = 10,000	4	0.002
	5	0.002
	6	0.003

Table 6. Simulation Parameters and Results for R_Dgas

R_Dgas	Number Collocated Devices	Pr{Major Source}
X_bar = 0.216	1	0.000
$s_X = 0.0203$	2	0.000
$\nu = n - 1 = 16$	3	0.000
iterations = 10,000	4	0.000
(X reflects normality	5	0.000
rather than lognormality)	6	0.000